NOVEL VIEW VIDEO PREDICTION USING A DUAL REPRESENTATION

Center for Research In Computer Vision
University of Central Florida

Sarah Shiraz
sarah.shiraz@knights.ucf.edu

Krishna Regmi
krishna.regmi7@gmail.com

Shruti Vyas
shruti@crcv.ucf.edu

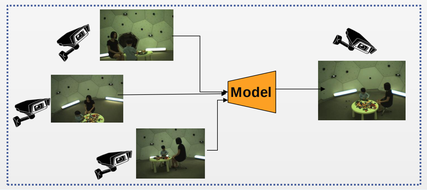
Yogesh S Rawat
yogesh@crcv.ucf.edu

Mubarak Shah
shah@crcv.ucf.edu

ICIP 2021
19-22 September 2021 • Anchorage, Alaska, USA

Problem Statement

We address the problem of novel view video prediction; given a set of input video clips from a single/multiple viewpoints, our network is able to predict the video from a novel viewpoint.



Contribution

- The proposed network integrates both global as well as view-dependent representation for effective novel view video prediction.
- The proposed framework can be generalized for varying number of input views and trained on a large number of views where we demonstrated its effectiveness for 72 different views.
- We provide extensive evaluation of our approach on two real-world datasets, achieving significant improvement over the existing methods.
- The proposed approach does not require any priors and is able to predict the video from wider angular distances, upto 45 degree, as compared to most of the recent studies predicting small variations in viewpoint.
- Our method relies only on RGB frames and camera parameters to learn a dual representation which is used to generate the video from a novel viewpoint
- We report 26.1% improvement in SSIM, 13.6% improvement in PSNR and 60% improvement in FVD scores over state-of-the-art viewLSTM on CMU-Panoptic Dataset.

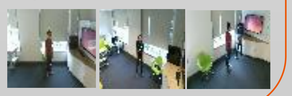
Dataset

CMUPanoptic
- Large scale real world dataset
- 480 VGA cameras, 31 HD cameras, and 10 Kinects
- 15 different activities of social interactions
- Upto 8 persons per activity
- Split of 6,244 training and 1,132 test samples

Samples

NTURGBD
- 60 action classes
- 3 camera viewpoints
- 40 participants
- 56,880 video clips
- cross subject evaluation split - 40,320 and 16,560 clips for training and testing respectively

Samples



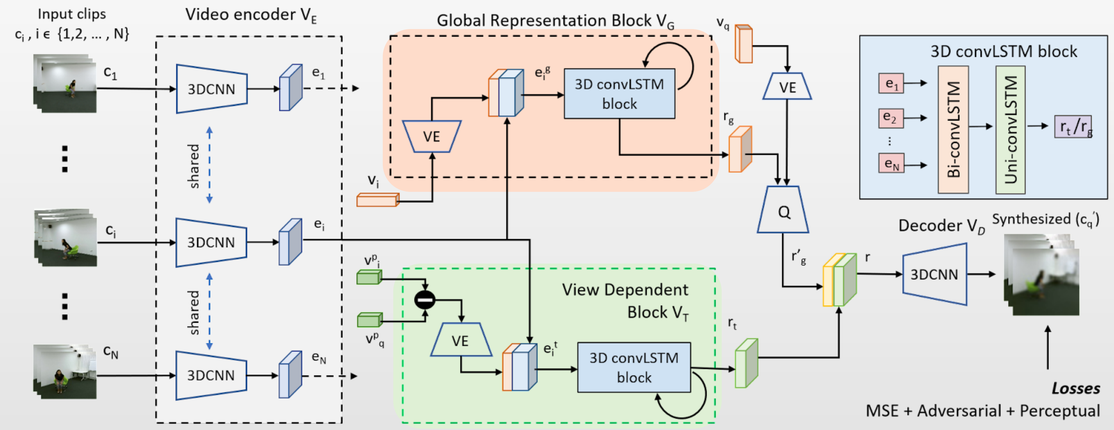
Related Work

ViewLSTM (Lakhal et al)
- Requires prior information from target view
- Combines input video features and prior (pose or depth) features using a novel recurrent structure i.e. viewLSTM

Image based approaches
- Transformable Bottleneck Networks (ICCV 2019)
- Multi-view Supervision for Single-view Reconstruction via Differentiable Ray Consistency (CVPR 2017)
- Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations (NIPS 2019)

Methodology

Network Architecture



- **Input**: set of N video clips from single/multiple viewpoints along with the respective view information
- **Output**: a synthesized video clip from a novel viewpoint.
- The N clips are of T×W×H×C dimension each, where T is number of frames, W width, H height and C number of channels. The view information is in the form of a vector, v of dimension N ×d where d is the number of view parameters for a viewpoint.
- A deep neural network which learns a dual internal representation, r , by assimilating global and view dependent information required for a novel view generation

Video Encoder
- Multiview encoder used to learn features of each clip in parallel.
- Uses a modified version of I3D architecture
- Shares weights for all the input video clips

Global Representation Block
- Transforms the features taken from a set of video clips according to a novel view.
- Outcome is a global representation representing the dynamics of the overall scene.
- Employs recurrent convolutional network to combine features from all input viewpoints

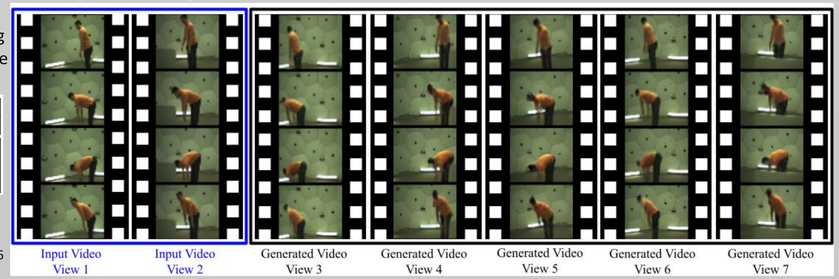
View Dependent Block
- Learns the features specific to the query view, helps in adding finer details
- More attention is assigned to the features from nearby/similar views to the query view.
- uses the input view information as well as query view information, in order to learn view dependent features

Experiments & Results

- In this setup, we use 72 camera viewpoints (CMUPanoptic Dataset).
- Testing is done by fixing 5 input views and using the remaining viewpoints one by one as query view.
- The scores affirm that the proposed method is successful at synthesizing high quality video clips from multiple query views

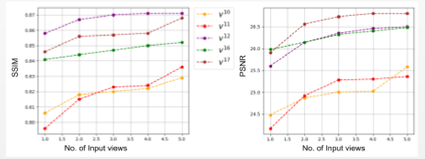
Panels	Metrics		
	SSIM(↑)	PSNR(↑)	FVD(↓)
4	0.837	25.15	11.52
5	0.848	24.9	13.42
17	0.845	25.46	13.31

Quantitative results in terms of SSIM, PSNR and FVD scores for Large-scale Multiview experiment (72 views), averaged over 19 query views of each of the three panels of CMU Panoptic Dataset at 56 × 56 resolution.



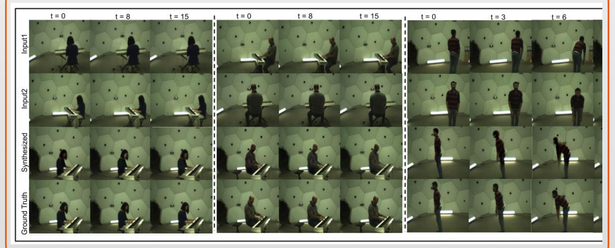
Experiments & Results

Effect of number of input views on video quality



SSIM and PSNR scores by varying the number of input views showing higher number of views help better prediction quality. The plots show scores for 5 randomly chosen query views. With increase in the number of input views the video quality improves.

Qualitative Results on Higher Resolution Videos



Qualitative Results on CMU Panoptic dataset with multi-view input setup showing three samples and the resolution is 224×224

Qualitative Results on NTU RGBD Dataset

Network Ablations for evaluating the role of the global and view-dependent representations

Networks	Metrics		
	SSIM(↑)	PSNR(↑)	FVD(↓)
VD (tested)	0.335	12.58	18.93
GR (tested)	0.639	17.45	15.68
VD (trained)	0.563	16.22	16.77
GR (trained)	0.792	20.05	13.45
Full	0.817	20.77	12.31

SSIM, PSNR and FVD scores for Network Ablations on NTU-RGB+D Dataset using View Dependent (VD) stream, Global Representation (GR) stream and the Full model (Full) on 56 × 56 resolution with multi-view input setup. The table shows the average scores for all input-output view combinations.